AD_____

Award Number: DAMD17-03-1-0047

TITLE: Therapy Selection by Gene Profiling

PRINCIPAL INVESTIGATOR: Simon W. Hayward, Ph.D.

CONTRACTING ORGANIZATION: Vanderbilt University
Nashville, TN 37232-2765

REPORT DATE: April 2006

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01-04-2006	Annual	1 Apr 2005 – 31 Mar 2006

4. TITLE AND SUBTITLE

Therapy Selection by Gene Profiling

5a. CONTRACT NUMBER

5b. GRANT NUMBER
DAMD17-03-1-0047

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Simon W. Hayward, Ph.D.

Email: simon.hayward@vanderbilt.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Vanderbilt University
Nashville, TN 37232-2765

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The long term goal of this work is to develop a new prognostic tool with which to determine the response of a patient to a given therapy, with the view of providing the most appropriate treatments tailored to individual patients. The central hypothesis of this proposal is that a subset of the genes expressed in a prostate tumor can be used to predict response to specific therapeutic regimens. The purpose of this work is to generate predictive methods which will allow patients to be selected for specific treatment protocols. In this year, per our proposed schedule, we have continued to focus on acquisition of tissue samples and their grafting and treatment in SCID mouse hosts. Collection and treatment of tissues is now completed. All samples have been assessed for response to Taxotere. Preparation of RNA is complete and microarrays have been run. 12 microarrays are being repeated to confirm results. A preliminary biostatistical analysis has been performed and a full analysis is underway.

15. SUBJECT TERMS
Taxotere, Genomics, Pharomacogenomics, microarrays

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	12	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Annual Report

PCRP Idea Development Award

DAMD 17-03-1-0047

Therapy selection by gene profiling

P.I. Simon W. Hayward, PhD

Introduction

The **long-term goal** of this work is to develop a new prognostic tool with which to determine the response of a patient to a given therapy, with the view of providing the most appropriate treatments tailored to individual patients. The **central hypothesis** of this proposal is that a subset of the genes expressed in a prostate tumor can be used to predict response to specific therapeutic regimens. The **purpose** of this work is to generate predictive methods that will allow patients to be selected for specific treatment protocols. A subordinate aim is to catalogue genes that are regulated in response to treatment with Taxotere, in both responding and non-responding human prostate cancer tissue samples, since these genes might suggest additional targets for therapeutic intervention. The **rationale** is to utilize a novel method of tissue grafting in combination with state of the art microarray, biostatistical and bioanalytical analysis to generate new prognostic tools. This project is an essential "proof of principle" step in the sense that if this methodology is successful with Taxotere it should be applicable to any new therapeutic approach that exists or which will be developed in the future. This project is divided into four specific aims. The first three aims will primarily generate data on the predictive value of gene expression profiles in samples derived from patients in determining the response of those patients to treatment with Taxotere. This work will allow us to design new predictive microarray or multiplex real time reverse transcriptase polymerase chain reaction (RT-PCR) assays to determine whether specific patients will respond to Taxotere. The fourth specific aim will use existing samples from patients engaged in an ongoing clinical trial to test whether these predictions are valid in a clinical setting.

Original Statement of Work and Current State of Progress

Therapy selection by gene profiling.

Task 1

Generate deoxyribose nucleic acid (DNA) Microarray patterns for prostate cancer samples from 150 patient tumor samples

a) As cases present, collect 150 histopathologically-confirmed prostate cancer tissue cores. Snap freeze core fragments (months 1-24) **Current status: Modified per first annual report because of changes in clinical practice to grafting of samples into untreated mice (also for Task 1b). Completed per modification.**

b) Prepare ribose nucleic acid (RNA) from snap frozen core fragments (months 1-25) **Current status: Completed per modification noted in 1a.**

c) Run 150 comparative DNA microarrays using 12k human chip against a mixed sample human prostate standard (months 1-25). **Current status: Modified to use 30k oligonucleotide chips. Completed.**

Task 2.

In vivo studies

a) Perform preliminary study to determine optimal post-treatment time point for determining histopathological response to Taxotere (months 1-3). **Current status: Completed.**

b) Graft tissues from the cores used in task 1 to pairs of severe combined immune-deficient (SCID) mice (months 1-24). **Current status: Completed.**

c) After 30 days treat one of each pair of mice with Taxotere for 6 days (months 1-25). **Current status: Completed.**

d) Sacrifice mice and harvest tissues. Snap freeze tissues, make RNA for microarray analysis, take representative tissue samples for histology (months 2-26). **Current status: Completed.**

e) Run 300 comparative microarrays of untreated vs standard and Taxotere-treated vs standard samples (months 3-28). **Current Status: Completed (Proviso: a small number of additional arrays are being run to confirm previous results.)**

Task 3
Biostatistical analysis

a) Identification of gene expression patterns which predict histopathologic response to Taxotere. Biostatistical analysis to determine a pattern of gene expression in tissue cores which predict histopathologic response to Taxotere in a xenograft model (months 26-32). **Current Status: Ongoing.**

b) Identification of genes regulated by Taxotere in responsive and non-responsive tissues. Biostatistical and bioinformatic analysis will be used to identify genes regulated by Taxotere in responsive and non-responsive tissue samples (months 28-34). **Current Status: Ongoing.**

c) Design of assays microarrays to predict response to Taxotere. Custom microarrays or assays (depending upon the number of prognostic genes identified in 3a) will be designed in which expression patterns of a limited number of genes should predict the response of human prostate cancer to Taxotere (months 32-35). **Current Status: Pending .**

Task 4
Prediction of response of patients in a clinical trial setting (months 35-36) **Current Status: Not yet initiated.**

Based upon microarray analysis of archived snap frozen tissue the ability of the arrays designed in task 3 to predict response in a clinical trial will be tested. Prepare RNA from archived tissue cores. Run microarrays, predict response based upon data acquired in earlier tasks. Test results by breaking patient code and correlating actual and predicted responses.

Work Ongoing and Completed

This report is a revision of the third annual report which clarifies progress through year three and into a no cost extension period. This revision is generated in response to a request to clarify an apparent lack of progress in the third annual report. As noted in our request for a no cost extension for this project, progress was limited in years two and three by the slow rate of collection of viable prostate cancer patient samples. This reflected two factors, the first being the migration towards robotic surgery which negatively impacts the quality of tissue recovered, this was noted in the first annual report. We have also noticed a slow but steady decrease in the size of the tumors in patients who are undergoing surgery. This resulted in a significant decrease in the frequency of availability of patient samples, since tissues only become available for research after standard of care pathology has been performed. As a result the collection period to reach our target number of samples was significantly longer than initially anticipated.

The third annual report reflected the progress on the grant at the time which was essentially collecting and processing samples. This was interpreted as a lack of progress. This appearance did not match reality, in that samples continued to be collected and grafted into animals, the animals were treated with Taxotere as appropriate and samples were collected. However given the linear nature of the project it was not possible to move onto the next stage until the samples were all available. These aspects were described when we requested that this grant be moved into a no cost extension. The prescribed number of samples was achieved soon after the submission of the third annual report. Since then progress has been rapid. All samples have been amplified using techniques which were validated early in the study. Patient samples responding to the drug treatment have been identified. The vast majority of microarrays have been run. A few (12) repeats are ongoing in relation to task 2e. Other than this tasks one and two have been successfully completed.

A preliminary biostatistical analysis has been undertaken and is described briefly below. **These details are provided as a basic reference since essentially the same methods are being used to assess gene expression profile differences in Taxotere responder versus non-responder groups and will also be applied to determine genetic changes in response to Taxotere treatment**. Specific gene expression profiles will not be listed to avoid premature disclosure in a public document of information with the potential for commercial applications.

The only potential problem identified so far is that the proportion of samples responding to Taxotere was lower than the predicted value of around 50%. The observed value was slightly below 20%. It is unknown at this time whether this will negatively affect the analysis, however the preliminary review from the biostatisticians involved (who have analyzed many similar studies) is that the sample numbers are sufficiently high that this should not be a problem.

The initial endpoint tested was differences in gene expression between samples in testosterone supplemented and castrated hosts. These experiments were performed to test whether Taxotere is more effective in patients who have or have not been subject to androgen ablation. The vast bulk of responders were in the androgen ablated group, which is the situation in which patients are most likely to receive the drug.

The data were imported from the microarray core in the form of .gpr files. These were used to create a single object of class "marrayRaw", this step was performed using the read.GenePix method of R package marray to generate a box plot of the raw ratio (cy5/cy3) array data. Since the raw data box plot showed variation among arrays, optimized local intensity-dependent normalization (OLIN) was applied. This step used the olin method of R package OLIN.

The four steps to select significant genes between cast and +Testosterone were:
1) Calculate paired t statistics between cast and +T for each gene. This step used the t.test method in R package stats. When one of paired data is missing, both of the paired data for this patient of the specific gene were set to NA. When the length of paired non-NA values was less than 2, the t value for this gene was set to 0.
2) Calculate local false discovery rate for each gene based on paired t values. This step used the locfdr method of R package.
3) Set up locfdr cutoff as 0.10, keep the genes that had locfdr <= 0.10 for further analyses. This step identified 475 genes.
4) Set up locfdr cutoff as 0.01 and NA% less than 40%, this reduced the candidate size to 58 significant genes. These 58 genes included obvious candidates such as PSA. This result provides confidence that the basic method works and will provide useful data to identify genes regulated by Taxotere in this model.

Further biostatistical manipulations were performed to identify pathways and groups of genes coordinately regulated in this system. Some details of the methods used are presented here because these will be applied in the analysis of genes regulated in Taxotere responders and non-responders.

Imputation of missing data based on nearest neighbor averaging (KNN)

For the top 470 genes the locfdr were <0.10, and the NA% were less than 40%. Since the data set had some NA missing data, imputation had to be performed before any multivariate data analyses. The nearest neighbor averaging (KNN) approach was applied. This step used the impute.knn method of the R impute package, the default number of neighbors to be used in the imputation was 10. The imputed data set with 470 genes were used in all of the following analyses.

Significant gene selection and GEE (Generalized Estimating Equation) modeling

The regression model suitable for this data set was Generalized Estimating Equation (GEE). Two methods were used to assemble group gene information before GEE modeling:

1) Principal component analysis (PCA) and GEE modeling

Principal component analysis (PCA) is a traditional method to reduce the number of dimensions of a multivariate data set. Three steps were used in this section:

Step 1: PCA: PCA was conducted on the imputed 470 gene data sets. Since the number of genes was greater than the sample size traditional R-mode PCA will not work. Therefore Q-mode PCA was applied. This step was performed using the prcomp method of R package stats.

Step 2: Data preparation: Since the contribution of the first 22 principal components was the first cutoff to achieve > 85% contribution, 22 PCs were used for GEE. However 22 features were too much for GEE modeling, therefore in a first examination only the first 5 PCs were used. These contributed 64.52% of the variation of 470 genes. To validate GEE modeling, a randomly selected data set with half of the patients was used as a training data set, and the data of the remaining samples were used as testing data set.

Step 3: GEE modeling on the first 5 PCs of the training data set was conducted by the geese method of the R package geepack, exchangeable correlation matrix was used as the correlation structure.

The accuracy of GEE modeling on top 5 PCs was 100% for the training data set, and 75% for the testing data set. The accuracy levels themselves were not low, but the problem with this approach was how to interpret each of the principal components. When we looked at the loading of these top 5 PCs, no significant contributing genes were found. In fact many of genes contributed a little for each principal component.

2) KEGG pathway analysis and GEE modeling

To overcome the interpretation problem in PCA and GEE analysis, instead of using principal components, KEGG pathways were applied. Three steps were used in this section:

Step 1: KEGG enrichment: Based on the 470 genes, LLD (LocusLink ids) were identified, and then KEGG pathway ids for the unique LLD of these 470 genes' were identified. For genes that have multiple LLDs, only the first LLD was taken for coding purposes. We used multiple methods in R package KEGG to conduct this step. Frequencies for each pathway id were calculated, the pathways were selected when their frequencies were at least 2 from 470 genes. Then from human pathway mapping, 5 out of 7 pathways had 5 human genes. Since GEE modeling can only take at most 5 features if we keep the same training data set as in PCA and GEE modeling, only the top 5 pathways that had most human genes were used for further analysis.

Step 2: Pathway score calculation: to get pathway score for each sample, we used paired t statistic as weight to calculate weighted linear combination of the involved genes for each pathway, The idea to calculate pathway score was from WFCCM., the formula was:

Score for pathway i and sample j = $\sum$k (t values of gene k) * (normalized ratio of gene k in pathway i and sample j).

Step 3: Data preparation: same training and testing data sets in PCA and GEE modeling section were applied here.

Step 4: GEE modeling on the first 5 KEGG pathways of the training data set was also conducted by geese method of R package geepack, exchangeable correlation matrix was again used as the correlation structure.

The accuracy of GEE modeling on the top 5 KEGG pathways was 93.75% for the training data set, and 81.25% for the testing data set. Comparing with PCA and GEE modeling results the overall accuracies were very close. In fact, the pathway approach had even higher accuracy for testing training than the PCA approach while their accuracies for training data set were comparable. Furthermore, the advantage of using the pathway approach was that it was easy to interpret results and the KEGG pathways are the standard biological pathways that are very meaningful in biology.

Classification analysis assuming independence

All of above analyses are for paired data when samples were not independent. Currently, there are not many methods available for paired data. Assuming independence among samples and using the same training and testing data sets as GEE part, we tried two classification methods based on 58 significant genes that passed paired t test when locfdr was less than 0.01 and NA% was less than 40%:

1) Support vector machine (SVM) with linear kernel and C-classification and default setting, this step used the svm of R package e1071, the top 58 significant genes showed very good classification accuracies in training and testing data sets:

2) Supervised clustering: performs supervised clustering of predictor variables, works in a greedy forward strategy and optimizes a combination of the Wilcoxon and Margin statistics for finding clusters. This step used the method wilma of R package supclust.

Since these classifications were applied to the 58 significant genes that passed paired t test, and which were suitable for paired data, our independent assumption was made only on the classification, our approaches for svm and supervised clustering were better than methods that totally ignored paired data construction.

Technical modifications:

Technical modifications relating to changes in clinical practice which has affected tumor tissue collection and changes in the microarray format from that originally proposed were noted in the first annual report. No significant changes have occurred in the current period.

Personnel Changes

None

Key Research Accomplishments

- Samples collected and processed through animals.
- RNA prepared and amplified.
- Microarray analysis completed.
- Response outcomes determined.
- Initial biostatistical analysis completed, full analysis ongoing.

Reportable Outcomes.

None

Conclusions.

This work is behind its timeline as a result in unavoidable delays in collecting patient samples. These delays reflected both changes in clinical practice (common adoption of robotic surgery) and earlier staging of patients resulting in lower chances of obtaining material from any given patient. Despite these obstacles samples have been collected and processed as required although at a delayed rate. RNA has been prepared, amplified and microarrays have been run. Biostatistical analysis of the resulting data is ongoing and should soon be completed. Completion of this stage will allow us to move forwards with task 3c and task 4.